Statistics and Probability Letters 89 (2014) 1–7

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Nonparametric estimation of cumulative incidence functions for competing risks data with missing cause of failure[☆]

Georgios Effraimidis[a,b,*], Christian M. Dahl[a,b,c]

[a] Department of Business and Economics, University of Southern Denmark, Denmark
[b] COHERE, University of Southern Denmark, Denmark
[c] CREATES, University of Southern Denmark, Denmark

ARTICLE INFO

Article history:
Received 7 February 2013
Received in revised form 1 February 2014
Accepted 3 February 2014
Available online 15 February 2014

Keywords:
Cumulative incidence function
Inverse probability weighting
Kernel estimation
Local linear estimation
Martingale central limit theorem

ABSTRACT

In this paper, we develop a fully nonparametric approach for the estimation of the cumulative incidence function with Missing At Random right-censored competing risks data. We obtain results on the pointwise asymptotic normality as well as the uniform convergence rate of the proposed nonparametric estimator. A simulation study that serves two purposes is provided. First, it illustrates in detail how to implement our proposed nonparametric estimator. Second, it facilitates a comparison of the nonparametric estimator to a parametric counterpart based on the estimator of Lu and Liang (2008). The simulation results are generally very encouraging.

1. Introduction

Competing risks models are widely used in biostatistics, empirical health economics and labor economics, for example, when analyzing data for onset of diseases, mortality due to mutually exclusive causes of death or unemployment where each individual is faced with competing exits (full-time employment, part-time employment). Hence, studying estimators of the cumulative incidence function within this modeling framework is of great importance. The goal of this paper is to derive asymptotic results for the nonparametric estimator of the cumulative incidence function in cases where continuous covariates affect the realization of the failure time and the cause of failure is Missing At Random (MAR) for some observations. The proposed nonparametric estimator is complementary to (i) the developed (semi)-parametric procedures with right-censored data and continuous explanatory covariates (e.g., Andersen et al., 1993; Jeong and Fine, 2007; Scheike et al., 2008 and (ii) the suggested parametric methods for right-censored data where the cause of failure is sometimes missing (Lu and Liang, 2008). Finally, we compare our results on uniform convergence rates with the results of Bordes and Gneyou (2011) who discuss the uniform convergence rate for the nonparametric estimator with right-censored competing risks data.

☆ We thank Gerard J. van den Berg, Christoph Breunig, Enno Mammen, and Geert Ridder for insightful comments and suggestions. We are also grateful to the two referees for a number of suggestions which helped us to improve the presentation. Finally, we would like to thank Wenbin Lu for sharing his code with us. All errors are our own responsibility. The second author acknowledges support from CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

* Correspondence to: Department of Business and Economics, Campusvej 55, 5230 Odense M, Denmark.
E-mail addresses: georgiosnl@gmail.com, gef@sam.sdu.dk (G. Effraimidis).

http://dx.doi.org/10.1016/j.spl.2014.02.001
0167-7152/

2. The nonparametric estimator

For expositional convenience, we will focus on two risks, 1 and 2. Let Y be the (actual) failure time and $\gamma \in \{1, 2\}$ be a failure type indicator. Let $X \in \mathcal{X} \subset \mathcal{R}^d$ be a vector of observed covariates and denote by x its realization. Define for each risk $j = 1, 2$ and $(t, x) \in \mathcal{R}_+ \times \mathcal{X}$ the cumulative incidence function

$$F_j(t|x) := \mathbb{P}(Y \leq t, \gamma = j|x). \tag{1}$$

We introduce the cause-specific hazard rate

$$\lambda_j(t, x) := \lim_{t\to 0} \frac{\mathbb{P}(t \leq Y < t + dt, \gamma = j|Y \geq t, x)}{t}. \tag{2}$$

The cumulative cause-specific hazard rate is defined as follows: $\Lambda_j(t, x) := \int_0^t \lambda_j(u, x)du$. Also, consider the overall hazard rate $\lambda(t, x) := \lambda_1(t, x)+\lambda_2(t, x)$, the corresponding cumulative overall hazard rate $\Lambda(t, x) := \int_0^t \lambda(u, x)du$, and the survival function $S(t - |x) := \mathbb{P}(Y \geq t|x)$. By using (1) and (2) we get for $j = 1, 2$

$$F_j(t|x) = \int_0^t S(u - |x)d\Lambda_j(u, x), \tag{3}$$

where

$$S(t - |x) = \prod_{u<t} \{1 - d\Lambda(u, x)\}. \tag{4}$$

Denote by Z the censoring variable with $Z \perp\!\!\!\perp Y, \gamma \mid X$, where the symbol $\perp\!\!\!\perp$ implies independence between the underlying random variables. Also, $T := \min(Y, Z)$, $\tilde{\gamma} := \gamma 1\{Y \leq Z\}$. We observe n independently and identically distributed copies $(T_i, X_i, 1\{\tilde{\gamma}_i > 0\}, R_i, R_i\tilde{\gamma}_i,)$, where R_i is the missing indicator variable and the missing data mechanism satisfies the MAR assumption (Rubin, 1976; Little and Rubin, 1987). The value of R_i equals 0 if $T_i = Y_i$ and the cause of failure is not observed. On the other hand, the indicator variable R_i is equal to 1 if either $T_i = Y_i$ and the cause of failure is observed or $T_i = Z_i$. The MAR scheme that we adopt is described as follows:

$$\mathbb{P}(R = 1|\tilde{\gamma}, \tilde{\gamma} > 0, T, x) = \mathbb{P}(R = 1|\tilde{\gamma} > 0, x) =: \pi(x). \tag{5}$$

The independence of the probability on T has as its consequence the predictability of all integrands of the proposed estimator. We also assume that

$$R \perp\!\!\!\perp T|X. \tag{6}$$

The latter is necessary in order to ensure that the underlying martingale processes are zero-mean. In the above discussion we assume that the covariates are time-invariant. This setup is adopted only for notational convenience as all the results in the sequel are true if X is predictable.

We will study the two following estimators for the cumulative incidence function,

$$\hat{F}_j^C(t|x) = \int_0^t \hat{S}^C(u - |x)d\hat{\Lambda}_j^C(u, x), \quad j = 1, 2 \tag{7}$$

and

$$\hat{F}_j^L(t|x) = \int_0^t \hat{S}^L(u - |x)d\hat{\Lambda}_j^L(u, x), \quad j = 1, 2, \tag{8}$$

where the superscripts C and L refer to the type of smoothing with respect to vector x. In particular, C is used for the local constant smoothing, whereas L is used for the local linear smoothing.

Let $\omega = (\omega_1, \dots, \omega_d) \in \mathcal{R}^d$ and $\mathcal{K}_h(\omega) = \frac{1}{h^d}\prod_{p=1}^d K\left(\frac{\omega_p}{h}\right)$, where K is a kernel with compact support $\mathbb{K}$ and $h = o(n)$. Introduce the quantity $\mathcal{L}_{h,x}(\omega) = \frac{\mathcal{K}_h(\omega)-\mathcal{K}_h(\omega)\omega^{\mathrm{T}}\bar{D}^{-1}\bar{c}_1}{\bar{c}_0-\bar{c}_1^{\mathrm{T}}\bar{D}^{-1}\bar{c}_1}$, with $\bar{c}_0 = \frac{1}{n}\sum_{i=1}^n \mathcal{K}_h(x - X_i)$, $\bar{c}_{1\rho} = \frac{1}{n}\sum_{i=1}^n \mathcal{K}_h(x - X_i)(x_\rho - X_{i\rho})$, $\bar{d}_{\rho\kappa} = \frac{1}{n}\sum_{i=1}^n \mathcal{K}_h(x - X_i)(x_\rho - X_{i\rho})(x_\kappa - X_{i\kappa})$, $\bar{c}_1 = (\bar{c}_{1\rho})_{\rho=1}^d$, and $\bar{D} = (\bar{d}_{\rho\kappa})_{\rho,k=1}^d$. The notations x_ρ and $X_{i\rho}$ refer to the ρ-th element of the corresponding row vector. The quantity $\mathcal{K}_h(.)$ will be used for the construction of the weights for the local constant estimator. On the other hand, the quantity $\mathcal{L}_{h,x}(.)$, which is also commonly referred to as the equivalent kernel, will be used for the construction of the weights for the local linear estimator.

First, we will describe the nonparametric estimator for the probability of having an observation with a missing cause of failure. The estimator of this probability is needed for the estimator of $\Lambda_j(t, x)$. Define $\pi(x, \tilde{\gamma}) := \mathbb{P}(R = 1|x, \tilde{\gamma}) = 1\{\tilde{\gamma} > 0\}\pi(x) + 1\{\tilde{\gamma} = 0\}$. That is, $\pi(x, \tilde{\gamma})$ specifies the probability of having an observation with a missing cause of failure given the observed characteristics x and the value of the indicator $\tilde{\gamma}$. This probability is independent of the exact value of $\tilde{\gamma}$ in case the latter is strictly positive (i.e., 1 or 2), whereas it is equal to one if $\tilde{\gamma} > 0$ (i.e., the observation is censored). For the local constant smoothing we have $\hat{\pi}^C(x) = \frac{\sum_{i=1}^n \mathcal{K}_h(x-X_i)1\{\tilde{\gamma}_i>0\}R_i}{\sum_{i=1}^n \mathcal{K}_h(x-X_i)1\{\tilde{\gamma}_i>0\}}$, whereas for the local linear smoothing we have $\hat{\pi}^L(x) = \frac{\sum_{i=1}^n \mathcal{L}_{h,x}(x-X_i)1\{\tilde{\gamma}_i>0\}R_i}{\sum_{i=1}^n \mathcal{L}_{h,x}(x-X_i)1\{\tilde{\gamma}_i>0\}}$.